Inferring from an imprecise Plackett–Luce model: application to label ranking

Loïc Adam, Arthur Van Camp, Sébastien Destercke, and Benjamin Quost

UMR CNRS 7253 Heudiasyc, Sorbonne Université, Université de Technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France
loic.adam@etu.utc.fr, {arthur.van-camp, sebastien.destercke, benjamin.quost}@hds.utc.fr

Abstract. Learning ranking models is a difficult task, in which data may be scarce and cautious predictions desirable. To address such issues, we explore the extension of the popular parametric probabilistic Plackett–Luce model, often used to model rankings, to the imprecise setting where estimated parameters are set-valued. In particular, we study how to achieve cautious or conservative inference with it, and illustrate their application on label ranking problems, a specific supervised learning task.

Keywords: Preference learning · Cautious inference · Poor data

1 Introduction

Learning and estimating probabilistic models over rankings of objects has received attention for a long time: earlier works can be traced back at least to the 1920s [21]. Recently, this problem has known a revival, in particular due to the rising interest of machine learning in the issue [12]. Popular approaches range from associating a random utility to each object to be ranked, from which a distribution on rankings is derived [3], to directly defining a parametric distribution over the set of rankings [19].

Multiple reasons motivate making cautious inferences of ranking models. The information at hand may be scarce — this is typically the case in the cold-start problem of a recommender system, or partial — for instance because partial rankings are observed (e.g., pairwise comparisons, top-k items). In addition, since inferring a ranking model is difficult and therefore prone to uncertainty, it may be useful to output partial rankings as predictions, thus abstaining to predict when information is unreliable.

Imprecise probability theory is a mathematical framework where partial estimates are formalized in the form of sets of probability distributions. Therefore, it is well suited to making cautious inferences and address the aforementioned problems; yet, to our knowledge, it has not yet been applied to ranking models.

In this paper, we use the imprecise probabilistic framework to infer a imprecise Plackett–Luce model, which is a specific parametric model over rankings,

from data. We present the model in Section 2. We address its inference in Section 3, showing that for this specific parametric model, efficient methods can be developed to make cautious inferences based on sets of parameters. Section 4 will then present a direct application to label ranking, where we will use relative likelihoods [5] to proceed with imprecise model estimation.

2 Imprecise Plackett–Luce models

In this paper, we consider the problem of estimating a probabilistic ranking model over a set of objects or labels $\Lambda = \{\lambda_1, \dots, \lambda_n\}$. This model defines probabilities over *total orders on the labels*—that is, complete, transitive, and asymmetric relations $\succ$ on Λ. Any complete order $\succ$ over the labels can be identified with its induced permutation or *label ranking* τ, that is the unique permutation of Λ such that

$$\lambda_{\tau(1)} \succ \lambda_{\tau(2)} \succ \cdots \succ \lambda_{\tau(n)}.$$

We will use the terms "order on the labels", "ranking" and "permutation" interchangeably. We denote by $\mathcal{L}(\Lambda)$ all $n!$ permutations on Λ, and denote a generic permutation by τ.

We focus on the particular probability model $P\colon 2^{\mathcal{L}(\Lambda)} \to [0,1]$ known as the Plackett–Luce (PL) model [6, 13]. It is parametrised by n parameters or *strengths* $v_1, \dots, v_n$ in $\mathbb{R}_{>0} \coloneqq \{x \in \mathbb{R} : x > 0\}$.[1] The *strength vector* $v = (v_1, \dots, v_n)$ completely specifies the PL model. For any such vector, an arbitrary ranking τ in $\mathcal{L}$ is assigned probability

$$P_v(\tau) \coloneqq \prod_{k=1}^{n} \frac{v_{\tau(k)}}{\sum_{\ell=k}^{n} v_{\tau(\ell)}} = \frac{v_{\tau(1)}}{v_{\tau(1)} + \cdots + v_{\tau(n)}} \cdot \frac{v_{\tau(2)}}{v_{\tau(2)} + \cdots + v_{\tau(n)}} \cdots \frac{v_{\tau(n-1)}}{v_{\tau(n-1)} + v_{\tau(n)}}. \tag{1}$$

Clearly, the parameters $v_1, \dots, v_n$ are defined up to a common positive multiplicative constant, so it is customary to assume that $\sum_{k=1}^{n} v_k = 1$. Therefore, the parameter $v = (v_1, \dots, v_n)$ can be regarded as an element of the interior of the n-simplex $\Sigma \coloneqq \{(x_1, \dots, x_n) \in \mathbb{R}^n_{\geq 0} : \sum_{k=1}^{n} x_k = 1\}$, denoted $\operatorname{int}(\Sigma)$.

This model has the following nice interpretation: the larger a weight v_i is, the more preferred is the label λ_i. The probability that λ_i is ranked first is

$$\sum_{\substack{\tau \in \mathcal{L}(\Lambda) \\ \tau(1) = \lambda_i}} P_v(\tau) = v_i;$$

conditioning on λ_i being the first label, the probability that λ_j is ranked second (i.e. first among the remaining labels) is equal to $v_j / \sum_{k=1, k \neq i}^{n} v_k$. This reasoning can be repeated for each of the labels in a ranking. As a consequence, given a PL model defined by v, finding the "best" (most probable) ranking amounts to finding the permutation $\tau^\star_v$ which ranks the strengths in decreasing order:

$$\tau^\star_v \in \operatorname*{arg\,max}_{\tau \in \mathcal{L}(\Lambda)} P_v(\tau') \Leftrightarrow v_{\tau(1)} \geq v_{\tau(2)} \geq v_{\tau(3)} \cdots \geq v_{\tau(n-1)} \geq v_{\tau(n)}. \tag{2}$$

[1] We also define the set of non-negative real numbers as $\mathbb{R}_{\geq 0} \coloneqq \{x \in \mathbb{R} : x \geq 0\}$.

We obtain an *imprecise* Plackett–Luce (IPL) model by letting the strengths vary over a subset Θ of $\text{int}(\Sigma)$ [2]. Based on this subset of admissible strengths, we can compute the *lower* and *upper probabilities* of a ranking τ as

$$\underline{P}_\Theta(\tau) \coloneqq \inf_{v \in \Theta} P_v(\tau) \quad \text{and} \quad \overline{P}_\Theta(\tau) \coloneqq \sup_{v \in \Theta} P_v(\tau) \quad \text{for all } \tau \text{ in } \mathcal{L}(\Lambda).$$

The above notion of "best" ranking becomes ambiguous for an IPL model, since two vectors $v \neq u \in \Theta$ might be associated with different "best" rankings $\tau_v^\star \neq \tau_u^\star$.

Therefore, we consider two common ways to extend (2). The first one, *(Walley–Sen) maximality* [22, 23], considers that τ_1 dominates τ_2 (noted $\tau_1 \succ_M \tau_2$) if it is more probable for any $v \in \Theta$:

$$\tau_1 \succ_M \tau_2 \Leftrightarrow (\forall v \in \Theta),\, P_v(\tau_1) > P_v(\tau_2). \tag{3}$$

The set $\mathcal{M}_\Theta$ of maximal rankings is composed of all such undominated rankings:

$$\mathcal{M}_\Theta \coloneqq \{\tau \in \mathcal{L}(\Lambda) : \nexists \tau' \text{ s.t. } \tau' \succ_M \tau\}. \tag{4}$$

We may have $|\mathcal{M}_\Theta| > 1$ when Θ is imprecise.

The second one is *E-admissibility* [18]. A ranking τ is *E-admissible* if it is the "best", according to Equation (2), for some $v \in \Theta$. The set $\mathcal{E}_\Theta$ of all E-admissible rankings is then

$$\mathcal{E}_\Theta \coloneqq \bigcup_{v \in \Theta} \underset{\tau \in \mathcal{L}(\Lambda)}{\arg\max}\, P_v(\tau) = \left\{\tau : (\exists v \in \Theta) \text{ s. t. } (\forall \tau' \in \mathcal{L}(\Lambda)),\, P_v(\tau) \geq P_v(\tau')\right\}. \tag{5}$$

By comparing Equations (4) and (5), we immediately find that $\mathcal{E}_\Theta \subseteq \mathcal{M}_\Theta$.

3 Learning an imprecise Plackett–Luce model

We introduce here two methods for inferring an IPL model. The first one (Section 3.1), which does not make further assumptions about Θ, provides an outer approximation of the set of all maximal rankings. The second one (Section 3.2) computes the set of E-admissible rankings via an exact and efficient algorithm, provided that the set of strengths Θ has the form of *probability intervals*.

3.1 General case

Section 2 shows that the "best" ranking is found using Equation (2). In the case of an IPL model, making robust and imprecise predictions requires to compare all possible ranks in a pairwise way, the complexity of which is $n!$—and thus generally infeasible in practice. However, checking maximality can be simplified. Notice that the numerator in Equation (1) does not depend on τ (product terms

[2] Taking $\text{int}(\Sigma)$ rather than Σ assures that all probabilities are positive and that Equation (1) is well-defined.

can be arranged in any order). Hence, when comparing two permutations τ and τ' using Equation (3), only denominators matter: indeed, $\tau \succ \tau'$ iff for all $v \in \Theta$,

$$\frac{P_v(\tau)}{P_v(\tau')} = \frac{v_{\tau'(1)} + \cdots + v_{\tau'(n)}}{v_{\tau(1)} + \cdots + v_{\tau(n)}} \cdot \frac{v_{\tau'(2)} + \cdots + v_{\tau'(n)}}{v_{\tau(2)} + \cdots + v_{\tau(n)}} \cdots \frac{v_{\tau'(n-1)} + v_{\tau'(n)}}{v_{\tau(n-1)} + v_{\tau(n)}} > 1. \quad (6)$$

Assume for a moment that strengths are precisely known, and that τ and τ' only differ by a swapping of two elements: $\tau(k) = \tau'(k)$ for all $k \in \{1, \ldots, m\} \backslash \{i, j\}$ where $i \neq j$, and $\tau(j) = \tau'(i)$, $\tau(i) = \tau'(j)$. Assume, without loss of generality, that $i < j$. Then, the product terms in Equation (6) only differ in the ratios involving rank j but not rank i; using furthermore $\tau(i) = \tau'(j)$, we get

$$\frac{P_v(\tau)}{P_v(\tau')} = \prod_{\substack{k=1 \\ k \notin \{i+1,\ldots,j\}}}^{n} \underbrace{\frac{\sum_{\ell=k}^{n} v_{\tau'(\ell)}}{\sum_{\ell=k}^{n} v_{\tau(\ell)}}}_{=1} \cdot \prod_{k=i+1}^{j} \frac{\sum_{\ell=k}^{n} v_{\tau'(\ell)}}{\sum_{\ell=k}^{n} v_{\tau(\ell)}} = \prod_{k=i+1}^{j} \frac{v_{\tau(i)} + \sum_{\ell=k,\ell\neq j}^{n} v_{\tau'(\ell)}}{v_{\tau(j)} + \sum_{\ell=k,\ell\neq j}^{n} v_{\tau(\ell)}}.$$

In this last ratio, we introduce now for any k in $\{i+1, \ldots, j\}$ the sums of strengths $C_k \coloneqq \sum_{\ell=k,\ell\neq j}^{n} v_{\tau(\ell)} = \sum_{\ell=k,\ell\neq j}^{n} v_{\tau'(\ell)}$: these terms being positive, it follows that

$$\tau \succ \tau' \quad \Leftrightarrow \quad \frac{P_v(\tau)}{P_v(\tau')} > 1 \quad \Leftrightarrow \quad (\forall v \in \Theta),\ v_{\tau(i)} > v_{\tau(j)}.$$

In the case of imprecisely known strengths, the latter inequality will hold whenever the following (sufficient, but not necessary) condition is met:

$$\underline{v}_{\tau(i)} \coloneqq \inf_{v\in\Theta} v_{\tau(i)} > \overline{v}_{\tau(j)} \coloneqq \sup_{v\in\Theta} v_{\tau(j)}.$$

Now comes a crucial insight. Assume a ranking τ which prefers λ_ℓ to λ_k whereas $\underline{v}_k > \overline{v}_\ell$, for some $k \neq \ell$: then, we can find a "better" ranking τ' (i.e., which dominates τ according to Equation (3)) by swapping labels λ_ℓ and λ_k. In other terms, as soon as $\underline{v}_k \geq \overline{v}_\ell$, all maximally admissible rankings satisfy $\lambda_k \succ \lambda_\ell$.

It follows that given an IPL model with strengths $\Theta \subseteq \mathrm{int}(\Sigma)$, we can deduce a partial ordering on objects from the pairwise comparisons of strength bounds: more particularly, we will infer that $\lambda_k \succ \lambda_\ell$ whenever $\underline{v}_k \geq \overline{v}_\ell$. This partial ordering can be obtained easily; it may contain solutions that are not optimal under the maximality criterion, but it is guaranteed to contain all maximal solutions.

3.2 Interval-valued case

We assume here that strengths are interval-valued: $v_k \in [\underline{v}_k, \overline{v}_k] \subseteq]0, 1[$; that is, the set Θ of possible strengths (called credal set hereafter) is defined by:

$$\Theta = \left(\bigtimes_{k=1}^{n} [\underline{v}_k, \overline{v}_k]\right) \cap \Sigma. \quad (7)$$

Note that we assume $\underline{v}_k > 0$ for each label λ_k: each object has a strictly positive lower probability of being ranked first. It follows that $\overline{v}_k < 1$, and thus $\Theta \subseteq \text{int}(\Sigma)$. Such interval-valued strengths fall within the category of *probability intervals on singletons* [1, Section 4.4], and are coherent (nonempty and convex) iff [10]:

$$(\forall k \in \{1,\dots,n\}),\ \Big(\underline{v}_k + \sum_{\substack{i=1\\ i\neq k}}^{n} \overline{v}_i \geq 1 \text{ and } \overline{v}_k + \sum_{\substack{i=1\\ i\neq k}}^{n} \underline{v}_i \leq 1\Big). \tag{8}$$

From now on, we will assume this condition to hold, and thus that Θ is coherent.

We are interested in computing the set of E-admissible rankings, i.e. rankings τ such that there exists $v \in \Theta$ for which τ maximises P_v (see Section 2). Our approach relies on two propositions, the proofs of which will be omitted due to the lack of place.

Checking E-admissibility We provide here an efficient way of checking whether a ranking τ is E-admissible. According to Equation (2), it will be the case iff v is decreasingly ordered wrt to τ, i.e. $v_{\tau(1)} \geq v_{\tau(2)} \geq v_{\tau(3)} \geq \dots$

Proposition 1. *Consider any interval-valued parametrisation of an IPL model such as defined by Equation* (7)*, and any ranking τ in $\mathcal{L}(\Lambda)$. Then, τ is E-admissible (i.e., $\tau \in \mathcal{E}_\Theta$) iff there exists an index $k \in \{1,\dots,n\}$ such that*

$$1 - \sum_{\ell=1}^{k-1} \min_{1\leq j\leq \ell} \overline{v}_{\tau(j)} - \sum_{\ell=k+1}^{n} \max_{\ell\leq j\leq n} \underline{v}_{\tau(j)} \quad \in \quad [\max_{k\leq j\leq n} \underline{v}_{\tau(j)}, \min_{1\leq j\leq k} \overline{v}_{\tau(j)}] \tag{9}$$

and

$$\begin{aligned} &\underline{v}_{\tau(\ell)} \leq \min\{\overline{v}_{\tau(1)},\dots,\overline{v}_{\tau(\ell)}\} && \text{for all } \ell \text{ in } \{1,\dots,k-1\},\\ &\overline{v}_{\tau(\ell)} \geq \max\{\underline{v}_{\tau(\ell)},\dots,\underline{v}_{\tau(n)}\} && \text{for all } \ell \text{ in } \{k+1,\dots,n\}. \end{aligned} \tag{10}$$

Checking E-admissibility via Proposition 1 has a polynomial complexity in the number n of labels. Indeed, we need to check n different values of k: for each one, Equation (9) requires to calculate a sum of $n-1$ terms, and Equation (10) to check $n-1$ inequalities, which yields a complexity of $2n(n-1)$.

Computing the set of E-admissible rankings Although Equation (9) opens the way to finding the set of E-admissible rankings, there are $n!$ many candidate rankings: checking all of them is intractable.

We propose to address this issue by considering a search tree, in which a node is associated with a specific sequence of labels. Each subsequent node adds a new element to this sequence: a leaf is reached when the sequence corresponds to a complete ranking. By navigating the tree top-down, we may progressively check whether a sequence corresponds to the beginning of an E-admissible ranking. Should it not, all completions of the sequence can be ignored.

This requires a way of checking whether a sequence $\kappa = (k_1, k_2, \dots, k_m)$, by essence incomplete, may be completed into an E-admissible ranking—i.e.,

whether we can find $\tau \in \mathcal{E}_\Theta$ such that $\tau(1) = k_1, \tau(2) = k_2, \ldots, \tau(m) = k_m$. Proposition 2 provides a set of necessary and sufficient conditions to this end.

Proposition 2. *Consider any coherent parametrisation of an IPL model such as defined by Equation* (7)*, and a sequence of distinct labels* $\kappa = (k_1, \ldots, k_m)$ *of length* $m \leq n-1$*. Then, there exists an E-admissible ranking beginning with this initial sequence iff the following equations are satisfied for every* j *in* $\{1, \ldots, m\}$*:*

$$\sum_{\ell=1}^{j} \min\{\overline{\nu}_{k_1}, \ldots, \overline{\nu}_{k_\ell}\} + \sum_{\substack{i=1 \\ i \notin \{k_1,\ldots,k_j\}}}^{n} \min\{\overline{\nu}_{k_1}, \ldots, \overline{\nu}_{k_j}, \overline{\nu}_i\} \geq 1; \tag{A_j}$$

$$\overline{\nu}_{k_j} \geq \max\{\underline{\nu}_i : i \in \{1, \ldots, n\} \setminus \kappa_j\}; \tag{B_j}$$

$$\sum_{t=0}^{j-1} \max\{\underline{\nu}_i : i \in \{1, \ldots, n\} \setminus \kappa_t\} + \sum_{\substack{i=1 \\ i \notin \{k_1,\ldots,k_j\}}}^{n} \underline{\nu}_i \leq 1; \tag{C_j}$$

here, κ_j *(*$j = 0, \ldots, m$*) is the sub-sequence of the* j *first labels in* κ *(by convention,* κ_0 *is empty), and* $\{1, \ldots, n\} \setminus \kappa_j$ *is the set of labels not appearing in* κ_j*.*

In the special case of $m = 1$, which is typically the case at depth one in the search tree, Equations (A_j), (B_j) and (C_j) reduce to:

$$\sum_{i=1}^{n} \min\{\overline{\nu}_{k_1}, \overline{\nu}_i\} \geq 1; \tag{A_1}$$

$$\overline{\nu}_{k_1} \geq \max\{\underline{\nu}_i : i \in \{1, \ldots, n\}\}; \tag{B_1}$$

$$\max\{\underline{\nu}_i : i \in \{1, \ldots, n\}\} + \sum_{\substack{i=1 \\ i \neq k_1}}^{n} \underline{\nu}_i \leq 1. \tag{C_1}$$

Note that under the coherence requirement (8), Equation (C_1) is a direct consequence of Equation (B_1), but it is not the case for Equation (C_j) when $j \geq 2$.

Example 1. Consider an IPL model that is defined by strength intervals $[\underline{\nu}_1, \overline{\nu}_1] = [3/8, 5/8]$, $[\underline{\nu}_2, \overline{\nu}_2] = [1/12, 1/12]$, $[\underline{\nu}_3, \overline{\nu}_3] = [1/30, 1/5]$ and $[\underline{\nu}_4, \overline{\nu}_4] = [1/8, 3/8]$, displayed in Figure 1 (the coherence of which can be checked using Equation (8)).

Consider the tree in Figure 2, which will help navigate the set of possible rankings with $n = 4$ labels. The left-most node at depth $m = 1$ corresponds to the sequence (λ_1); its left-most child (left-most node at depth $m = 2$) to the sequence (λ_1, λ_2). We can see that this sequence has been ruled out as a possible initial segment for an E-admissible ranking: no further completion (i.e., neither of the two rankings $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and $(\lambda_1, \lambda_2, \lambda_4, \lambda_3)$) will be checked.

The sequence $(\lambda_1, \lambda_3, \lambda_2)$ has been ruled out as well; however, the sequence $(\lambda_1, \lambda_3, \lambda_4)$ has been considered as valid, and can be straightforwardly completed

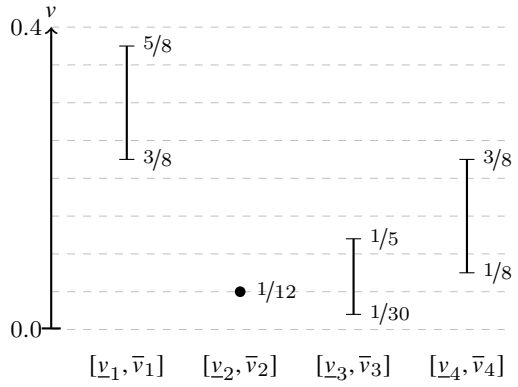

Fig. 1. Probability intervals for Example 1

into a valid E-admissible ranking (since only one possible label remains). Eventually, all E-admissible rankings $\tau = (\tau(1), \tau(2), \tau(3), \tau(4))$ corresponding to the IPL model are

$$\{(1,3,4,2), (1,4,2,3), (1,4,3,2), (4,1,3,2)\}.$$

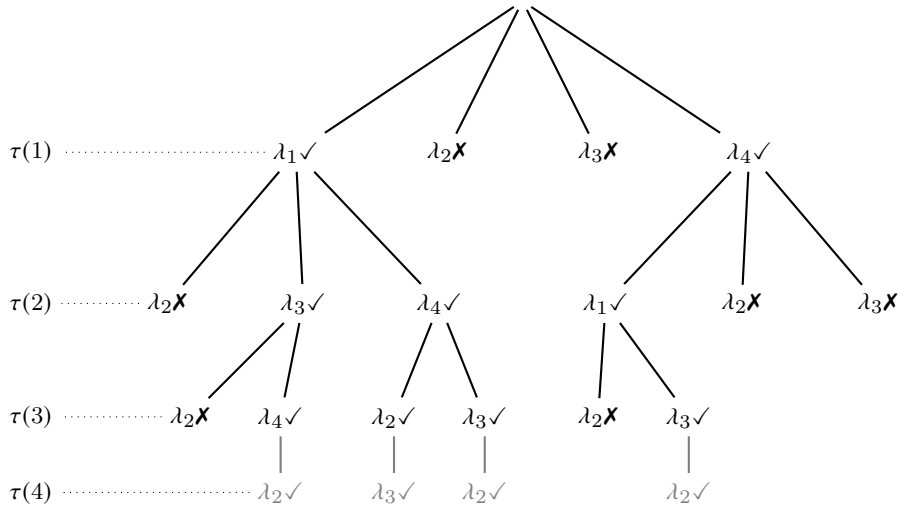

Fig. 2. Search tree for $n = 4$

A possible strength vector for which $\tau = (1, 3, 4, 2)$ dominates all others is given by $v = (5/8, 1/12, 1/6, 1/8)$: it can easily be checked that $v \in \Theta$ and that $v_{\tau(1)} = 5/8 \geq v_{\tau(2)} = 1/6 \geq v_{\tau(3)} = 1/8 \geq v_{\tau(4)} = 1/12$, i.e. τ is E-admissible according to Equation (2). We provide below possible strength vectors for each of the E-admissible rankings associated with the IPL model considered:

admissible strength vector $v \in \Theta$	corresponding ranking $\tau \in \mathcal{E}_\Theta$
$v = (v_1, v_2, v_3, v_4)$	$\tau = (\tau(1), \tau(2), \tau(3), \tau(4))$
$(5/8, 1/12, 1/6, 1/8)$	$(1, 3, 4, 2)$
$(5/8, 1/12, 1/12, 5/24)$	$(1, 4, 2, 3)$
$(5/8, 1/12, 1/12, 5/24)$	$(1, 4, 3, 2)$
$(3/8, 1/12, 1/6, 3/8)$	$(4, 1, 3, 2)$

Let us show that there is no E-admissible ranking τ that starts for instance with $(1, 2)$. Assume *ex absurdo* that such an E-admissible ranking τ exists. This would imply that there exists $v \in \Theta$ such that $v_1 \geq v_2 \geq \max\{v_3, v_4\}$, which by Equation (2) would imply that $1/12 = v_2 \geq v_4 \geq \underline{v}_4 = 1/8$, which is impossible. ◊

Algorithm Equations (A_j), (B_j) and (C_j) used in Proposition 2 to check the E-admissibility of a ranking with a given initial sequence of labels can be turned into an efficient algorithm. We can indeed proceed recursively: checking whether there exists an E-admissible ranking starting with $(k_1, \ldots, k_m)$ basically requires to check whether it is the case for $(k_1, \ldots, k_{m-1})$ and then whether Equations (A_j), (B_j) and (C_j) still hold for $j = m$.

Algorithms 1 and 2 provide a pseudo-code version of this procedure. Note that as all branch-and-bound techniques, it does not reduce the worst-case complexity of building an E-admissible set. Indeed, if all the rankings are E-admissible—which typically happens when all probability intervals are wide, then no single branch can be pruned from the search tree. In that case, the algorithm navigates the complete tree, which clearly has a factorial complexity in the number of labels n. Then, even a simple enumeration of all E-admissible rankings has such a complexity.

However, in practice we can expect many branches of the tree to be quickly pruned: indeed, as soon as one of the Equations (A_j), (B_j) or (C_j) fail to hold, a branch can be pruned from the tree. We expect this to allow for efficient inferences in many circumstances.

Algorithm 1 Find the E-admissible rankings opt_n

Require: Probability intervals $[\underline{v}_k, \overline{v}_k]$ for $k \in \{1, \ldots, n\}$
Ensure: The set $\Theta = \{[\underline{v}_k, \overline{v}_k] : k \in \{1, \ldots, n\}\}$ is coherent
  $\text{opt}_n \leftarrow \emptyset$
  for all $k_1 \in \{1, \ldots, n\}$ **do**
    `Recur`$(1, (k_1))$
  end for

Algorithm 2 `Recur`$(j, (k_1, \ldots, k_j))$

if $j = n - 1$ **then**
  append the unique $k_n \in \{1, \ldots, n\} \setminus \{k_1, \ldots, k_{n-1}\}$ to the end of $(k_1, \ldots, k_{n-1})$
  add $(k_1, \ldots, k_n)$ to opt_n ▷we found a solution.
else

  for all $k_{j+1} \in \{1, \ldots, n\} \setminus \{k_1, \ldots, k_j\}$ **do**

    if Equations (A_{j+1}), (B_{j+1}) and (C_{j+1}) hold **then**
      append k_{j+1} to the end of $(k_1, \ldots, k_j)$
      `Recur`$(j + 1, (k_1, \ldots, k_{j+1}))$
    end if
  end for
end if

4 An application to label ranking

In this section, we explore an application of the IPL model to supervised learning of label rankings. Usually, supervised learning consists in mapping any instance $\mathbf{x} \in \mathcal{X}$ to a single (preferred) label $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ representing its class. Here, we study a more complex issue called label ranking, which rather maps $\mathbf{x} \in \mathcal{X}$ to a predicted total order $\hat{y}$ on the labels in Λ—or a partial order, should we accept to make imprecise predictions for the sake of robustness.

For this purpose, we exploit a set of training instances associated with rankings $(\mathbf{x}_i, \tau_i)$, with $i \in \{1, \ldots, m\}$, in order to estimate the theoretical conditional probability measure $P_{\mathbf{x}} \colon 2^{\mathcal{L}(\Lambda)} \to [0, 1]$ associated to an instance $\mathbf{x} \in \mathcal{X}$. Ideally, observed outputs τ_i should be complete orders over Λ; however, this is seldom the case, total orders being more difficult to observe: training instances are therefore frequently associated with incomplete rankings τ_i (i.e., partial orders over Λ).

Here, we will apply the approach detailed in Section 3.1 to learning an IPL model from such training data, using the contour likelihood to get the parameter set corresponding to a specific instance $\mathbf{x}$.

4.1 Estimation and prediction

Precise predictions In [7], it was proposed to use an instance-based approach: the predictions for any $\mathbf{x} \in \mathcal{X}$ are made locally using its nearest neighbours.

Let $\mathcal{N}_K(\mathbf{x})$ stand for the set of nearest neighbours of $\mathbf{x}$ in the training set, each neighbour $\mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})$ being associated with a (possibly incomplete) ranking τ_i; and let M_i be the number of ranked labels in τ_i. Using the classical instance-based assumption that distributions are locally identical (i.e., in the neighborhood of $\mathbf{x}$), the probability of observing $\tau_1, \ldots, \tau_K$ given a parameter value v is:

$$P(\tau_1, \ldots, \tau_K | v) = \prod_{\mathbf{x}_i \in \mathcal{N}_K(\mathbf{x})} \prod_{m=1}^{M_i} \frac{v_{\tau_i(m)}}{\sum_{j=m}^{M_i} v_{\tau_i(j)}}. \tag{11}$$

We can then use maximum likelihood estimation (MLE) in order to determine v from $\tau_1, \ldots, \tau_K$, by maximizing (11)—or equivalently, its logarithm

$$l(v) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[\log(v_{\tau_i(m)}) - \log \sum_{j=m}^{M_i} v_{\tau_i(j)} \right].$$

Various ways to obtain this maximum have been investigated. We will use here the minorization-maximization (MM) algorithm [16], which aims, in each iteration, to maximize a function which minorizes the log-likelihood:

$$Q_k(v) = \sum_{i=1}^{K} \sum_{m=1}^{M_i} \left[\log(v_{\tau_i(m)}) - \frac{\log \sum_{j=m}^{M_i} v_{\tau_i(j)}}{\log \sum_{j=m}^{M_i} v^{(k)}_{\tau_i(j)}} \right]$$

where $v^{(k)}$ is the estimation of v in the k-th iteration. When the parameters are fixed, the maximization of Q_k can be solved analytically and the algorithm provably converges to the MLE estimate v^* of v. The best ranking τ^* is then

$$\tau^* \in \arg \max_{\tau \in \mathcal{L}(\Lambda)} P(\tau | v^*);$$

it is simply obtained by ordering the labels according to v^* (see Equation (2)).

Imprecise predictions An IPL model is in one-to-one correspondence with an imprecise parameter estimate, which can be obtained here by extending the classical likelihood to the contour likelihood method [5]. Given a parameter space Σ and a positive likelihood function L, the contour likelihood function is:

$$L^*(v) = \frac{L(v)}{\max_{v \in \Sigma} L(v)};$$

by definition, L^* takes values in $]0, 1]$: the closer $L^*(v)$ is to 1, the more likely v is. One can then naturally obtain imprecise estimates by considering "cuts". Given β in $[0, 1]$, the β-cut of the contour likelihood, written B^*_β, is defined by

$$B^*_\beta = \{v \in \Sigma : L^*(v) \geq \beta\}.$$

Once B^*_β is determined, for any test instance $\mathbf{x}$ to be processed, we can easily obtain an imprecise prediction $\hat{y}$ in the form of a partial ranking, using the results of Section 3.1: we will retrieve $\hat{y}$ such that $\lambda_i \succ \lambda_j$ for all $v_k \in B^*_\beta$. We stress here that the choice of β directly influences the precision (and thus the robustness) of the model: $B^*_1 = v^*$, which generally leads to a precise PL model; when β decreases, the IPL model is less and less precise, possibly leading to partial (and even empty) predictions.

In our experiments, the contour likelihood function is modelled by generating multiple strengths v according to a Dirichlet distribution with parameter $\beta = \gamma v^*$, where v^* is the ML estimate obtained with the best PL model (or equivalently, the best strength v) and $\gamma > 0$ is a coefficient which makes it possible to control the concentration of parameters generated around v^*.

4.2 Evaluation

When the observed and predicted rankings y and $\hat{y}$ are complete, various accuracy measures [15] have been proposed to measure how close they are to each other (0/1 accuracy, Spearman's rank, ...). Here, we retain Kendall's Tau:

$$A_\tau(y, \hat{y}) = \frac{C - D}{n(n-1)/2}, \tag{12}$$

where C and D are respectively the number of concording and discording pairs in y and $\hat{y}$. In the case of imprecise predictions $\hat{y}$, the usual quality measures can be decomposed into two components [9]: correctness (CR), measuring the accuracy of the predicted comparisons, and completeness (CP):

$$CR(y, \hat{y}) = \frac{C - D}{C + D} \quad \text{and} \quad CP(y, \hat{y}) = \frac{C + D}{n(n-1)/2}, \tag{13}$$

where C and D are the same as in Equation (12). Should $\hat{y}$ be complete, $C + D = n(n-1)/2$, $CR(y, \hat{y}) = A_\tau(y, \hat{y})$ and $CP(y, \hat{y}) = 1$; while $CR(y, \hat{y}) = 1$ and $CP(y, \hat{y}) = 0$ if $\hat{y}$ is empty (since no comparison is done).

4.3 Results

We performed our experiments on several data sets, mostly adapted from the classification setting [7]; we report here those obtained on the Bodyfat, Housing and Wisconsin data sets. For each dataset, we tested several numbers of neighbours: $K \in \{5, 10, 15, 20\}$ (for the MLE estimate and using Equation (12)), and chose the best by 10-fold cross-validation. The sets of parameters B^*_β were obtained as explained above, by generating 200 strengths with $\gamma \in \{1, 10\}$, the best value being selected via 10-Fold cross validation repeated 3 times.

We also compared our approach to another proposal [8] based on a rejection threshold of pairwise preference probabilities, in three different configurations:

- using the original, unperturbed rankings;
- by deleting some labels in the original rankings with a probability $p \in [0, 1]$;
- by introducing some noise in the rankings, by randomly swapping adjacent labels with a probability $p \in [0, 1]$ (the labels being chosen at random).

Figure 3 displays the results of both methods for the Bodyfat data set (with $m = 252$ and $n = 7$) when rankings remain unperturbed, with a confidence interval of 95% (±2 standard deviation of measured correctness). Our approach based on the contour likelihood function is on par with the method based on abstention, which was the case with all tested data sets. Both methods see correctness increase once we allow for abstention. On the other data sets, the same behaviour can be seen: our approach seems to be on par with the one based on abstention, provided that the contour likelihood function has been correctly modelled (i.e., the generation of strengths is appropriate).

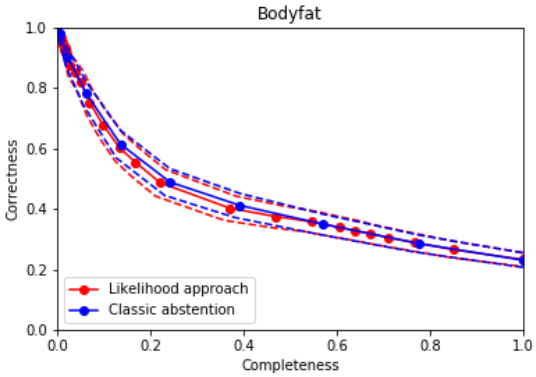

Fig. 3. Comparison of methods on Bodyfat with no perturbations

In order to be able to compare the two methods, we show underneath results on a specific range of the completeness. We only show the domain $[0.6, 1]$; however the behaviour is similar outside this range.

Figures 4 and 5 show that both methods are also on par on the Housing data set ($m = 506$, $n = 6$) even when the data sets are missing some labels. It can also be noticed that for a given completeness level, the correctness is lower than in the unperturbed case. On average, the greater the level of perturbation is, the lower the average correctness is. This also stands for the other data sets.

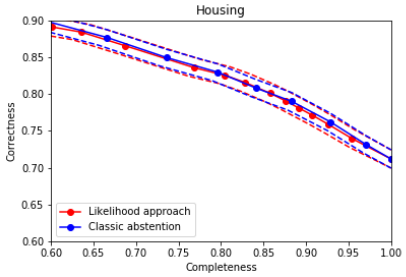

Fig. 4. Comparison of methods on Housing with no perturbations

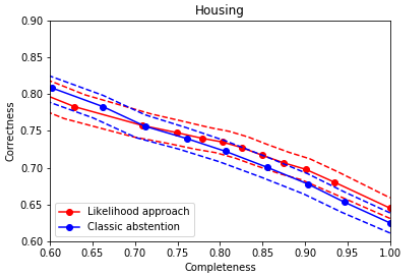

Fig. 5. Comparison of methods on Housing with 60% of missing label pairs

Figures 6 and 7 display that with a different method of perturbation (label swapping), our approach gives similar results on the Wisconsin data set ($m = 194$, $n = 16$). Moreover, the correctness is again lower in average for a given completeness level if the data set is perturbed. We observe the same behaviour for the label swapping perturbation method on the other data sets.

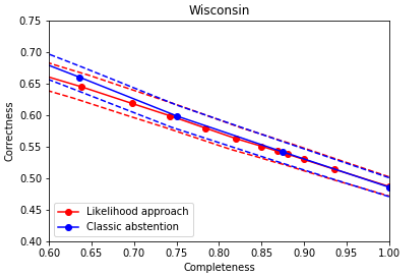

Fig. 6. Comparison of methods on Wisconsin with no perturbations

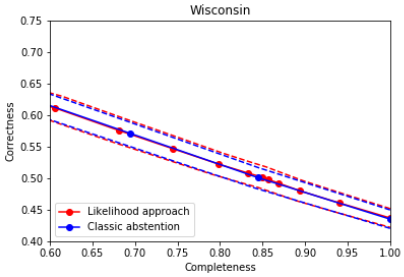

Fig. 7. Comparison of methods on Wisconsin with 60% of swapped label pairs

Such results are encouraging, as they show that we can at least achieve results similar to state-of-the-art approaches. We yet have to identify those cases where the two approaches significantly differ.

5 Conclusions

In this paper, we made a preliminary investigation into performing robust inference and making cautious predictions with the well known Plackett–Luce model, a popular ranking model in statistics. We have provided efficient methods to do so when the data at hand are poor, that is either of a low quality (noisy, partial) or scarce. We have demonstrated the interest of our approach in a label ranking problem, in presence of missing or noisy ranking information.

Possible future investigations may focus on the estimation problem, which may be improved, for example by extending Bayesian approaches [14] through the consideration of sets of prior distributions; or by developing a natively imprecise likelihood estimate, for instance by coupling recent estimation algorithms using stationary distribution of Markov chains [20] with recent works on imprecise Markov chains [17].

As suggested by an anonymous reviewer, it might be interesting to consider alternatives estimation methods such as epsilon contamination. There already exist non-parametric, decomposition-based approaches to label ranking with imprecise ranks; see [11, 4]. However, the PL model, being tied to an order representation, may not be well-suited to such an approach. We intend to investigate this in the future.

Last, since the Plackett–Luce model is known to be strongly linked to particular RUM models [2, 24], it may be interesting to investigate what becomes of this connection when the RUM model is imprecise (for instance, in our case, by considering Gumbel distributions specified with imprecise parameters).

Acknowledgements

This work benefited from the financial support of the projects PreServe ANR-18-CE23-0008 and LABEX MS2T ANR-11-IDEX-0004-02 of the French National Research Agency (ANR). We would like to thank three anonymous reviewers for their motivating comments.

Bibliography

[1] T. Augustin, F. P. Coolen, G. De Cooman, and M. C. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] H. Azari, D. Parks, and L. Xia. Random utility theory for social choice. In *Advances in Neural Information Processing Systems*, pages 126–134, 2012.

[3] G. Baltas and P. Doyle. Random utility models in marketing research: a survey. *Journal of Business Research*, 51(2):115–125, 2001.

[4] Y.-C. Carranza-Alarcon, S. Messoudi, and S. Destercke. Cautious label-wise ranking with constraint satisfaction. In M.-J. Lesot, S. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R. R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 96–111, Cham, 2020. Springer International Publishing.

[5] M. Cattaneo. *Statistical Decisions Based Directly on the Likelihood Function*. PhD thesis, ETH Zurich, 2007.

[6] W. Cheng, K. Dembczynski, and E. Hüllermeier. Label ranking methods based on the Plackett–Luce model. In *Proceedings of the 27th Annual International Conference on Machine Learning - ICML*, 2010.

[7] W. Cheng, E. Hüllermeier, and K. J. Dembczynski. Label ranking methods based on the plackett–luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222, 2010.

[8] W. Cheng, E. Hüllermeier, W. Waegeman, and V. Welker. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 25 (NIPS-12)*, pages 2510–2518, 2012.

[9] W. Cheng, M. Rademaker, B. De Baets, and E. Hüllermeier. Predicting partial orders: ranking with abstention. *Machine Learning and Knowledge Discovery in Databases*, pages 215–230, 2010.

[10] L. de Campos, J. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *I. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2:167–196, 1994.

[11] S. Destercke, M.-H. Masson, and M. Poss. Cautious label ranking with label-wise decomposition. *European Journal of Operational Research*, 246(3):927–935, Nov. 2015.

[12] J. Fürnkranz and E. Hüllermeier, editors. *Preference Learning*. Springer Berlin Heidelberg, 2011.

[13] J. Gu and G. Yin. Fast algorithm for generalized multinomial models with ranking data. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2445–2453, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[14] J. Guiver and E. Snelson. Bayesian inference for plackett–luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 377–384. ACM, 2009.

[15] E. Hüllermeier, J. Furnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172:1897–1916, 2008.

[16] D. R. Hunter et al. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

[17] T. Krak, J. De Bock, and A. Siebes. Imprecise continuous-time markov chains. *International Journal of Approximate Reasoning*, 88:452–528, 2017.

[18] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[19] J. Marden. *Analyzing and modeling rank data*, volume 64. Chapman & Hall/CRC, 1996.

[20] L. Maystre and M. Grossglauser. Fast and accurate inference of plackett–luce models. In *Advances in neural information processing systems*, pages 172–180, 2015.

[21] L. Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1927.

[22] M. Troffaes. Generalising the conjunction rule for aggregating conflicting expert opinions. *I. J. of Intelligent Systems*, 21(3):361–380, March 2006.

[23] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, New York, 1991.

[24] J. I. Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.